CRM 91-246 / February 1992

Reliability of Mechanical Maintenance Performance Measures

Paul W. Mayberry
William H. Wright

50 Years CNA 1992

CENTER FOR NAVAL ANALYSES
4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

Work conducted under contract N00014-91-C-0002.

REPORT DOCUMENTATION PAGE

Form Approved
OPM No. 0704-0188

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	February 1992	Final

4. TITLE AND SUBTITLE

Reliability of Mechanical Maintenance Performance Measures

5. FUNDING NUMBERS

C - N00014-91-C-0002

PE - 65153M

PR - C0031

6. AUTHOR(S)
Paul W. Mayberry, William H. Wright

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Center for Naval Analyses
4401 Ford Avenue
Alexandria, Virginia 22302-0268

8. PERFORMING ORGANIZATION REPORT NUMBER

CRM 91-246

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Commanding General
Marine Corps Combat Development Command (WF 13F)
Studies and Analyses Branch
Quantico, Virginia 22134

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

A fundamental requirement in the development and administration of performance measures is that such assessments should result in reliable scores that accurately indicate a person's level of proficiency. This research memorandum examines the reliability of two performance measures of mechanical maintenance developed for the Marine Corps Job Performance Measurement Project: hands-on performance tests and job knowledge tests. Multiple estimates of reliability were computed, and the consistency of test administrators in scoring hands-on performance was specifically examined.

The hands-on performance tests and the job knowledge tests were found to result in very reliable measurements. Properly trained and monitored test administrators were able to score hands-on performance consistently across examinees, over time, and for different test content. Implications for subsequent performance measurement are presented, and possible training implications based on mechanics who were retested are noted.

14. SUBJECT TERMS

Analysis of variance, JPM (job performance measurement), Maintenance personnel, Marine Corps personnel, Performance (Human), Performance tests, Proficiency, Reliability, Scoring, Statistical analysis, Test methods, Validation

15. NUMBER OF PAGES
50

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
CPR	CPR	CPR	SAR

50 Years CNA 1992 CENTER FOR NAVAL ANALYSES
4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

4 March 1992

MEMORANDUM FOR DISTRIBUTION LIST

Subj: CNA Research Memorandum 91-246

Encl: (1) CNA Research Memorandum 91-246, *Reliability of Mechanical Maintenance Performance Measures*, by Paul W. Mayberry and William H. Wright, Feb 1992

1. Enclosure (1) is forwarded as a matter of possible interest.

2. A fundamental requirement in the development and administration of performance measures is that such assessments should result in reliable scores that accurately indicate a person's level of proficiency. This research memorandum examines the reliability of two performance measures of mechanical maintenance developed for the Marine Corps Job Performance Measurement Project: hands-on performance tests and job knowledge tests. Multiple estimates of reliability were computed, and the consistency of test administrators in scoring hands-on performance was examined specifically.

3. The hands-on performance tests and the job knowledge tests were found to result in very reliable measurements. Properly trained and monitored test administrators were able to score hands-on performance consistently across examinees, over time, and for different test content. Implications for subsequent performance measurement are presented, and possible training implications based on mechanics who were retested are noted.

Lewis R. Cabe
Director
Manpower and Training Program

Distribution List:
Reverse page

Subj: Center for Naval Analyses Research Memorandum 91-246

Distribution List

SNDL

A1 DASN - MANPOWER (2 copies)
A1H ASSTSECNAV MRA
A2A CNR
A5 CHNAVPERS
A5 PERS-2
A5 PERS-5
A5 PERS-11
A6 CG MCRDAC - Washington
A6 HQMC AVN
A6 HQMC MPR & RA
  Attn: Code MR
  Attn: Code MP
  Attn: Code MM
  Attn: Code MA (3 copies)
  Attn: Code MPP-54
A6 HQMC PP&O
FF38 USNA
  Attn: Nimitz Library
FF42 NAVPGSCOL
FF44 NAVWARCOL
FJA1 COMNAVMILPERSCOM
FJA13 NAVPERSRANDCEN
  Attn: Technical Director (Code 01)
  Attn: Technical Library
  Attn: Dir., Manpower Systems (Code 11)
  Attn: Dir., Personnel Systems (Code 12)
FJB1 COMNAVCRUITCOM
FT1 CNET
V12 CG MAGTEC
V12 CG MCCDC
  Attn: Studies and Analyses Branch
  Attn: Director, Warfighting Center
  Attn: Warfighting Center, MAGTF Propenency and Requirements Branch (2 copies)
V12 CG MCRDAC - Quantico

OTHER

Defense Advisory Committee on Military Personnel Testing (8 copies)
Joint Service Job Performance Measurement Working Group (12 copies)
Military Accession Policy Working Group (17 copies)

CRM 91-246 / February 1992

Reliability of Mechanical Maintenance Performance Measures

Paul W. Mayberry
William H. Wright

Operations and Support Division

50 Years
CNA 1992
CENTER FOR NAVAL ANALYSES
4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

ABSTRACT

A fundamental requirement in the development and administration of performance measures is that such assessments should result in reliable scores that accurately indicate a person's level of proficiency. This research memorandum examines the reliability of two performance measures of mechanical maintenance developed for the Marine Corps Job Performance Measurement Project: hands-on performance tests and job knowledge tests. Multiple estimates of reliability were computed, and the consistency of test administrators in scoring hands-on performance was specifically examined.

The hands-on performance tests and the job knowledge tests were found to result in very reliable measurements. Properly trained and monitored test administrators were able to score hands-on performance consistently across examinees, over time, and for different test content. Implications for subsequent performance measurement are presented, and possible training implications based on mechanics who were retested are noted.

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

The joint-service Job Performance Measurement (JPM) Project was initiated to develop objective and reliable measures of job performance to be used in validating the military selection test. Hands-on performance tests and job knowledge tests were developed for this purpose. Before the scores from such performance measures can be meaningfully interpreted or applied, it is necessary to examine their consistency or reliability. The reliability of hands-on performance tests is of particular concern because of the dynamic between a test administrator and an examinee. This research memorandum examined the reliability of the performance measures administered to five mechanical maintenance military occupational specialties (MOSs): motor transport mechanic (MOS 3521) and four helicopter mechanic specialties (MOSs 6112, 6113, 6114, and 6115).

METHOD

Hands-on performance tests were constructed to represent the extensive range of mechanical maintenance job requirements. Hands-on tests composed of 15 to 20 tasks were constructed for each of the five MOSs. Written job knowledge tests were developed for each MOS to be parallel in content to its respective hands-on performance test. After thorough tryout, the final versions of the job knowledge tests included between 145 and 179 items. Automotive and helicopter mechanics were randomly selected for testing based on paygrade, time in service, and aptitude requirements so that the sample would be representative of their respective specialties. Over 1,000 automotive mechanics and over 150 mechanics for each helicopter type were tested for two days each.

Several estimates of reliability were computed for the performance measures of each MOS, as appropriate: test-retest reliability (compares performance on the same test given on two separate occasions), split-halves reliability (compares performance on separate halves of a test based on a single test administration), alpha coefficient (reflects the degree to which tasks or items measure the same concept), and scorer agreement (reflects consistency of performance scores assigned by two test administrators).

Another reliability procedure, called generalizability analysis, was used to estimate the variance in hands-on performance scores that is attributable to specific factors of the measurement process. Such factors should represent meaningful elements of the measurement process (differences in scores may vary by examinee, test administrator, testing site, performance task, etc.). Three measurement models were examined to assess the complex structure of hands-on tests.

RESULTS

Table I reports the various reliability estimates for each MOS. The different methods of computing reliability resulted in consistent and high estimates of reliability for both the hands-on and job knowledge tests. For the hands-on tests, the reliability estimates for the four helicopter MOSs were somewhat higher than those for the automotive mechanic MOS. Test-retest and split-halves reliabilities were equivalent, indicating that confidence can be placed in performance differences noted for the interval of 10 to 14 days. Split-halves reliabilities and alpha coefficients indicated that 15 to 20 tasks composing the hands-on tests measured related skills. High percent agreement was noted between test administrators when comparisons were aggregated over all tasks, test administrator pairs, and testing occasion. Similar results were noted for the job knowledge tests.

Table I. Reliability estimates for hands-on performance tests and job knowledge tests

	MOS				
Reliability measure	3521	6112	6113	6114	6115
Hands-on performance test					
Test-retest	.79	.88			
Split-halves	.80	.87	.91	.82	.87
Alpha coefficient	.81	.88	.88	.78	.81
Scorer agreement[a]	.96	.98	.96	.96	.94
Job knowledge test					
Test-retest	.73	.87			
Split-halves	.94	.97	.96	.95	.92
Alpha coefficient	.93	.97	.96	.94	.93

NOTE: Reliability estimates corrected for restrictions in range.

a. Percent agreement between two test administrators at the step level.

The measurement models from the generalizability analyses showed that test administrators consistently scored hands-on performance. Such scoring accuracy was consistent across individuals, for different test content, and across the full test administration period of five months. The majority of variance in hands-on scores was attributable to differences in performance by individual Marines. The 15 to 20 hands-on tasks were found to differ substantially in their level of difficulty. For the helicopter mechanics, substantial site effects of equivalent

magnitude and direction were noted for the same testing location and were found for each MOS. Further analyses of the helicopter data were conducted separately by site. Such site effects were not found for the automotive mechanics. A sample of mechanics was retested. Generalizability analysis of the retest data showed that examinees improved on retesting, they improved more on some tasks than others, and test administrators were consistent in scoring retest performance.

CONCLUSIONS

The assessment of performance via hands-on tests is potentially replete with factors that can degrade their reliability. The quality and consistency of scoring by test administrators is a primary concern. The analyses of this research memorandum have focused on estimating the reliability of the hands-on tests (and job knowledge tests) for several mechanical specialties so that policy-makers can understand the factors that affect such performance measurements.

Both the hands-on and job knowledge tests were found to result in reliable performance measurement based on several different methods of estimating reliability. Analyses of the factors that contribute to the variance of hands-on scores showed that the test administrators were interchangeable and had little impact on the overall hands-on scores. These findings were probably the result of several strategies employed during the test development and administration process intended to enhance the reliability of the measurements (e.g., thorough test tryout and revision, observable performance actions being the basis for scorable steps, and substantial test administrator training and monitoring).

Despite these efforts to obtain reliable measurements, hands-on test scores were found to vary by testing site for all of the helicopter MOSs. Detailed training materials were prepared, and the same team of trainers was used at both testing locations to standardize conditions and to minimize any arbitrary site differences. The presence of the site effect suggests that additional monitoring procedures are needed to check scoring consistency not only within test site but also across test sites.

The hands-on performance tests were composed of 15 to 20 tasks that were sampled to represent the extremely large domain of mechanical job requirements. The differences in task difficulty noted for each MOS indicate that selection of tasks may have significant implications for the interpretation of hands-on performance scores. Given such extensive job requirements for Marine Corps mechanics, certain aspects of the job domain may not be represented if the tasks are not properly sampled or if too few tasks are tested. It follows that explicit efforts should be made to sample tasks that are representative of the full range of job requirements.

The Marine Corps JPM Project has developed reliable measures of hands-on performance and job knowledge for several mechanical maintenance specialties. Confidence can be placed in the interpretation of the resulting scores, and conclusions of further analyses will not be affected by the instability or inconsistency of the performance scores. These reliability analyses also point to several potential implications for the training community, which will be addressed more completely in later analyses of the mechanical performance data.

CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

ILLUSTRATIONS

THIS PAGE INTENTIONALLY LEFT BLANK

TABLES

INTRODUCTION

All measurement contains some error. The magnitude of such error affects the ability to make confident statements about the concept being measured, so one attempts to design and administer measures that are reliable. The benefits of a reliable test are threefold. First, a reliable test will render consistent, meaningful measurements for the same examinees over time, given that the concept does not itself change. Conversely, if scores from a reliable test do change over time, one can reasonably assert that the change in scores reflects actual growth or deterioration and is not the result of random error. Second, the degree to which a test is reliable serves as a limit to its validity. In principle, a test cannot be more highly related to another test than it can to itself. Finally, the extent to which a test is reliable directly affects its generalization from the single measurement to a larger context. Therefore, the assessment of measurement reliability is an initial concern before any resulting scores can be meaningfully interpreted or applied.

The joint-service Job Performance Measurement (JPM) Project was initiated in the early 1980s to develop objective measures of job performance to be used in validating the military selection test. The measures that were developed for this purpose were hands-on performance tests and job knowledge tests. The hands-on tests required examinees to perform a sample of job tasks under realistic but standardized conditions, whereas the job knowledge tests covered items related to job performance that could not be tested in the hands-on mode. The Marine Corps JPM Project has focused on measuring performance for representative military occupational specialties (MOSs) within several of its larger occupational fields.

This research memorandum examines the reliability of the performance measures administered to five mechanical maintenance MOSs: motor transport mechanic (MOS 3521) and four helicopter mechanic specialties (CH-46, MOS 6112; CH-53A/D, MOS 6113; U/AH-1, MOS 6114; and CH-53E, MOS 6115). For each MOS, three topics related to reliability are addressed:

- Estimates of reliability for the hands-on and job knowledge tests
- Consistency of test administrators in scoring hands-on performance
- Factors of the measurement process that contribute to the variance of hands-on scores.

METHOD

Performance Measures

The hands-on performance tests were based on extensive job analyses and review of Marine Corps technical manuals and training materials to specify fully the domain of mechanical job requirements. Job task domains were developed for each MOS. Based on considerable input and review by job incumbents and experts in Marine Corps subject matter, tasks were sampled for testing to be representative of the broad and diverse requirements of each mechanical job. Extensive tryouts were conducted with job incumbents to construct a hands-on test for each sampled task so that its instructions were readily understood, it could be objectively scored, and it accurately reflected task performance on the job. Separate hands-on tests composed of 15 to 20 tasks were constructed for each of the five MOSs. Further details describing the hands-on development process are noted elsewhere [1].

Paper-and-pencil job knowledge tests were developed for each MOS to be parallel in content to its respective hands-on performance test. Development of the job knowledge tests began with a review of the steps of the hands-on tests that were crucial to the performance of the task. Items were then written to capture these critical aspects of task performance. To ensure performance-based written items, extensive use was made of graphic materials and illustrations. Additional test items were developed to cover content areas that could not be tested by the hands-on tests. The job knowledge tests were pilot-tested with a sample of job incumbents. The final versions of the tests included between 145 and 179 items and a time limit of about two hours.

Sample

Automotive and helicopter mechanics were stratified and then randomly selected for testing to satisfy paygrade, time in service, and aptitude requirements. The intent was that the tested sample would be representative of the respective populations of automotive and helicopter mechanics. Each Marine was tested for two days: one day of hands-on performance testing and a second day of job knowledge tests and other tests associated with predicting the performance measures.

Table 1 provides the descriptive statistics for the five MOSs that were tested. The automotive mechanics had a mean mechanical maintenance (MM) composite score[1] in the range of 110, and the helicopter mechanics had a mean MM score of about 115. The sample of mechanics tested covered a broad range of time in service, ranging from Marines just out of training to senior personnel with over ten years of service.

Table 1. Descriptive statistics for mechanical maintenance samples

	MOS				
Statistic	3521	6112	6113	6114	6115
Sample size	1,028	174	120	215	149
Enlisted MM aptitude score					
Mean	109.7	114.7	115.8	116.2	115.3
Standard deviation	11.4	9.1	11.2	8.7	10.9
Range	76-141	87-140	78-137	83-140	83-141
Time in service					
Mean	48.9	51.8	54.5	45.4	66.9
Standard deviation	36.9	38.3	45.5	31.7	44.7
Range	8-232	9-196	12-173	10-172	21-216

Estimates of Reliability

Several estimates of reliability were computed for both performance measures of each MOS, as appropriate:

- Test-retest reliability: comparison of performance on the same test given on two separate occasions. A sample of automotive and CH-46 mechanics was retested after an interval of 10 to 14 days.

1. The mechanical maintenance composite is used to determine which recruits are eligible for MOSs requiring mechanical aptitudes. In the 1980 national youth population, a representative sample of potential military applicants, the MM composite has a mean of 100 and a standard deviation of 20. The minimum MM scores to be eligible for the automotive and helicopter mechanic MOSs are 95 and 105, respectively. Waivers to these minimum aptitude requirements may be given in special cases.

- Split-halves reliability: estimation of measurement consistency based on a single administration of the test by comparing performance levels on separate halves of the test.

- Alpha coefficient: a measure of the internal consistency that reflects the degree to which hands-on tasks or test items measure the same concept.

- Scorer agreement: the percent agreement between two test administrators as they score the step-level performance of one mechanic.

These methods of estimating reliability reflect the impact of different sources of error on the performance measure.

Test-retest reliability is a measure of the "stability" of performance. The magnitude of a test-retest estimate will vary with changes in performance over time, the length of the time interval between test administrations, and measurement errors associated with both test administrations. Test-retest reliability is useful in determining the confidence with which generalizations can be made from a mechanic's score at one particular time to what he would obtain at a different time. However, changes in performance scores cannot be isolated solely as a function of true changes in the performance trait; other factors also affect the magnitude of the test-retest correlation.

Split-halves reliability, in combination with a test-retest estimate, provides further insight into interpreting changes in performance. By comparing the performance of mechanics on similar but different content, an estimate of "equivalence" is obtained. The split-halves procedure enables the analyst to determine how confidently a mechanic's score can be generalized to what he would obtain if he took a performance test composed of similar but different tasks or items. Any change in performance across the halves is considered error (sampling error due to differences in content across the halves). Because the halves were administered at the same time, there is no error due to change in the performance trait. However, estimates of split-halves reliability can be affected by the way in which the halves are formed and the quality of individual items in measuring the performance trait (as well as the number of items on the test).

Alpha coefficient, an estimate of the internal consistency of test items, reflects the homogeneity of test items by quantifying the degree to which item responses correlate with the total test score. If individual test items are not correlated, the reliability of the performance measure suffers because the test items do not consistently measure the same trait. Alpha coefficients are affected by both the correlations among all test items and the number of items on the test. In most circumstances, an alpha coefficient provides a conservative estimate of a performance measure's reliability.

For hands-on performance tests, the scoring by test administrators potentially introduces a unique source of error variance that is not explicitly measured by the previous reliability estimates. To the extent that test administrators deviate from their originally trained scoring standards or disagree over the successful performance of a task, the hands-on test will not be a reliable measure. Analyses were conducted to estimate the extent to which test administrators agreed in scoring of the performance of mechanics.

Generalizability Analyses

Each of the previous reliability measures yields a single estimate that describes the aggregate impact of multiple, undifferentiated sources of error. For example, a test-retest reliability coefficient confounds errors due to changes in the performance trait, inconsistent measurement due to differences in task difficulty, scoring errors by test administrators, and differences in administrator scoring patterns over time. It would be informative to identify the potential components of the measurement process that systematically contribute to the overall error of performance measurement. If the magnitude of such error sources can be estimated, policy-makers could be made aware of their influences on performance measurement or specific actions could be taken to reduce their impact in subsequent measurements.

Generalizability analysis (based on generalizability theory, or G-theory) allows for the simultaneous estimation of factors contributing to the variance of performance scores. Generalizability analysis is based primarily on analysis of variance (ANOVA) procedures that partition the variance of observed scores into separate components corresponding to main effects and their interactions. The factors included in the analysis should represent meaningful elements within the measurement process that possibly contribute to the overall variation of hands-on performance scores.

Estimating the impact of various factors of the measurement process on hands-on performance scores requires complex data collection designs that allow each factor to be modeled explicitly. From a measurement perspective, there are multiple sources of potential measurement error that can affect an estimate of a mechanic's performance level. Measurement error might arise from inconsistencies in test administrators, in task difficulty, or in administration differences on two different occasions. The magnitude of the variance component for such factors will be examined to determine the major contributors to variance in hands-on performance scores.

RESULTS

Reliability of Performance Measures

Table 2 reports the various reliability estimates for each MOS. The reliability estimates have been corrected for range restriction;

observed reliabilities and the calculations for the corrections are reported in appendix A. The various methods of computing reliability resulted in consistent and high estimates of reliability for both the hands-on and job knowledge tests. For the hands-on performance tests, the following observations were apparent from the results of table 2:

- Reliability estimates for the four helicopter MOSs were somewhat higher than those for the automotive mechanic MOS.
- Test-retest and split-halves reliabilities were equivalent, indicating stability of the mechanical performance trait over an interval of 10 to 14 days.
- Split-halves reliabilities and alpha coefficients indicated consistent measurement of the performance trait by the 15 to 20 tasks composing the hands-on tests.
- High percent agreement was noted between test administrators when comparisons were aggregated over all tasks, test administrator pairs, and testing occasion.

Table 2. Reliability estimates for hands-on performance and job knowledge tests

	MOS				
Reliability measure	3521	6112	6113	6114	6115
Hands-on performance test					
Test-retest	.79	.88			
Split-halves	.80	.87	.91	.82	.87
Alpha coefficient	.81	.88	.88	.78	.81
Scorer agreement[a]	.96	.98	.96	.96	.94
Job knowledge test					
Test-retest	.73	.87			
Split-halves	.94	.97	.96	.95	.92
Alpha coefficient	.93	.97	.96	.94	.93

NOTE: Reliability estimates corrected for restriction in range.

a. Percent agreement between two test administrators at the step level.

Similar results were noted for the job knowledge tests:

- Split-halves reliabilities and alpha coefficients were exceptionally high because of consistent measurement at the item level and the large number of items on each test (between 145 and 179 items).

- Test-retest estimates were slightly lower than the same estimates for the hands-on tests but were still sufficiently high to allow for confident generalizations over time.

The test-retest reliabilities for both tests were examined further to understand the gains in performance.

Figures 1 through 4 present the scatterplots of both tests for the retest versus initial scores. Mechanics from only two MOSs were retested: 88 automotive mechanics (MOS 3521) and 67 CH-46 mechanics (MOS 6112). The diagonal line in the figures indicates the line of no change; points above the line represent Marines who improved on retesting, and points below the line represent Marines whose performance went down on retesting.

Mean retest gains in hands-on performance were approximately .7 and .4 standard deviation for MOSs 3521 and 6112, respectively. Such improvement in performance may be the result of practice or a better understanding of the testing procedure. Hands-on retest gains were found to be negatively related to aptitude (the mechanical maintenance aptitude composite that is used by the Marine Corps to determine eligibility for mechanically related jobs). That is, lower-aptitude Marines improved their hands-on performance scores more than Marines of higher aptitude. This outcome resulted from the already high hands-on performance scores achieved by the higher-aptitude mechanics, so there were fewer opportunities for substantial improvement.

Similar mean retest gains of less than half a standard deviation were noted for the job knowledge tests. Figure 3 shows somewhat more scatter of the data points for the automotive mechanics, as was indicated by the slightly lower test-retest correlation found for this group. The job knowledge tests were moderately difficult; the average mechanic answered about 62 and 71 percent of the items correctly for the automotive and helicopter specialties, respectively. Gains in job knowledge scores were not as strongly related to aptitude as was found for the hands-on tests, but the trend was still that lower-aptitude Marines made larger gains.

Consistency of Test Administrator Scoring

Table 2 showed that the step-level agreement indexes of administrator pairs were consistently in the middle to upper 90-percent range for all MOSs. However, such aggregate comparisons potentially overlook

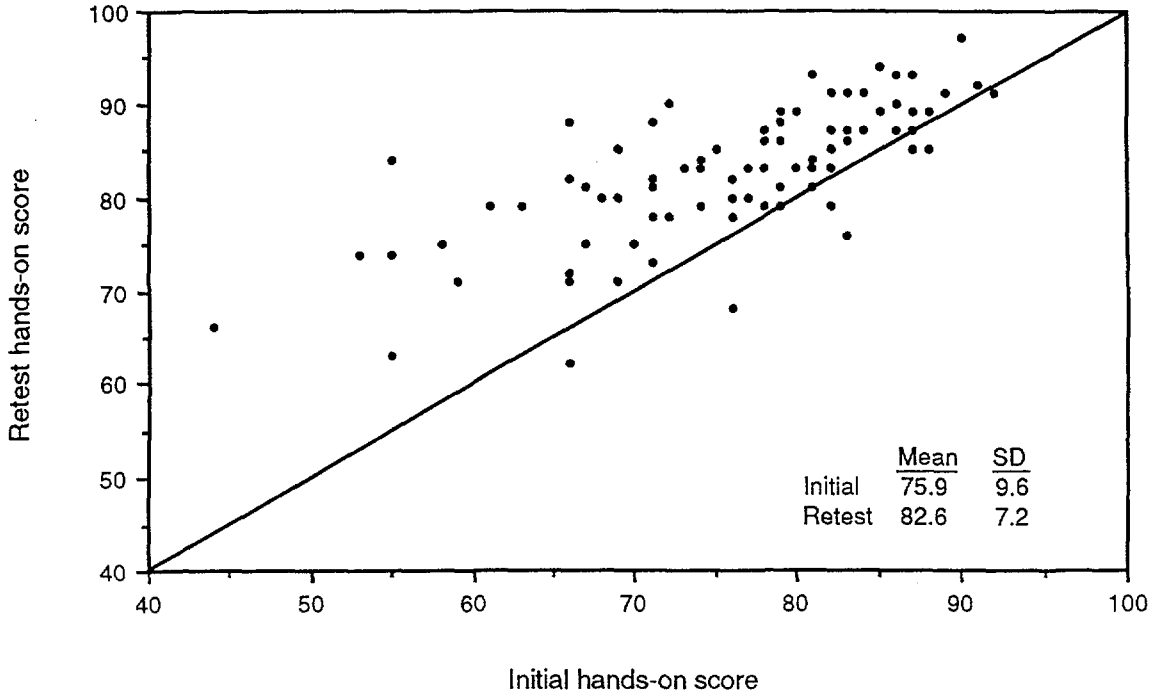


Figure 1. Scatterplot of initial and retest hands-on scores for ground mechanics (3521)

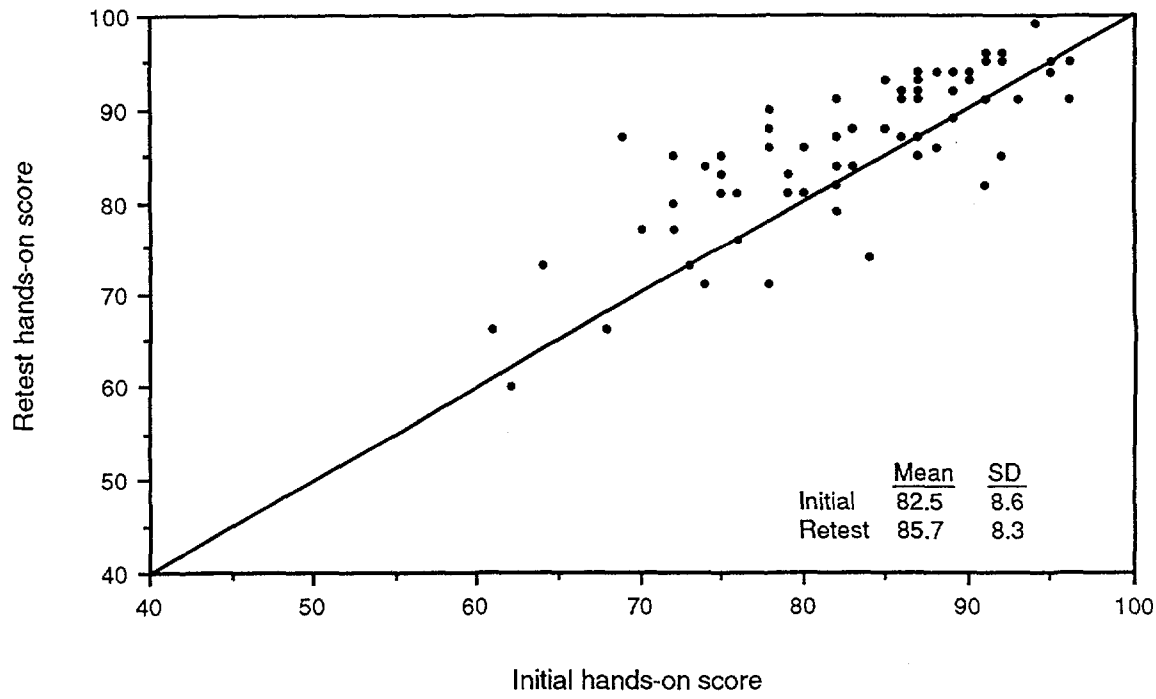


Figure 2. Scatterplot of initial and retest hands-on scores for CH-46 mechanics (6112)

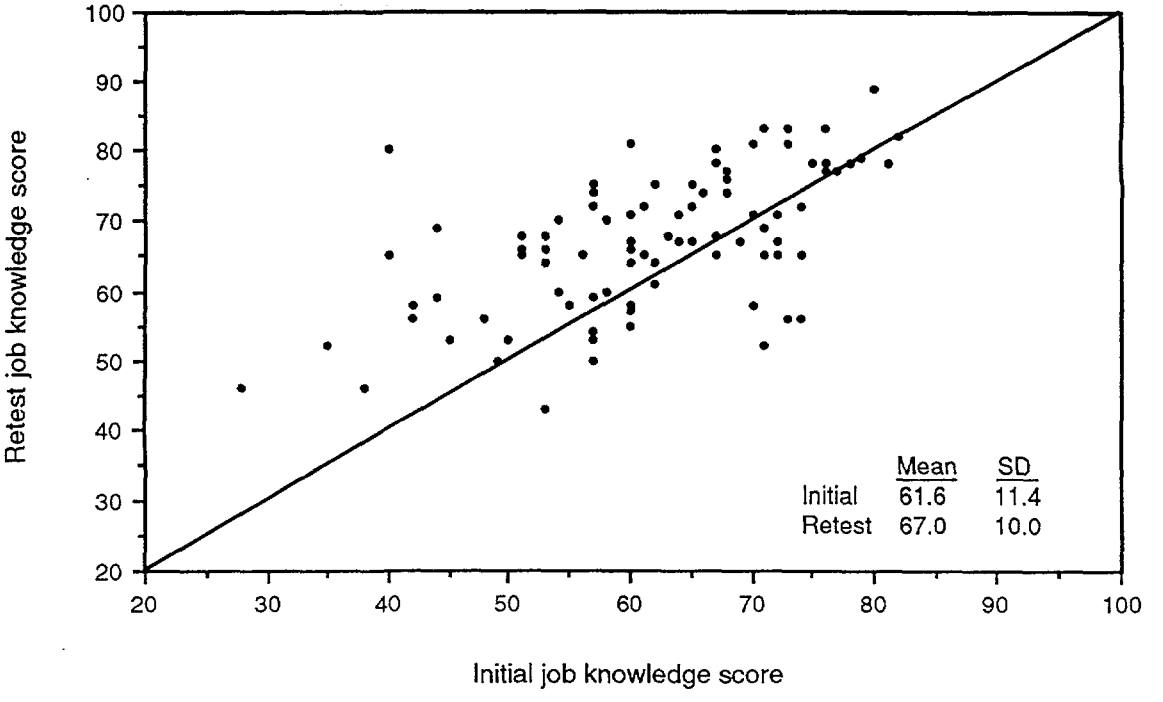


Figure 3. Scatterplot of initial and retest job knowledge scores for ground mechanics (3521)

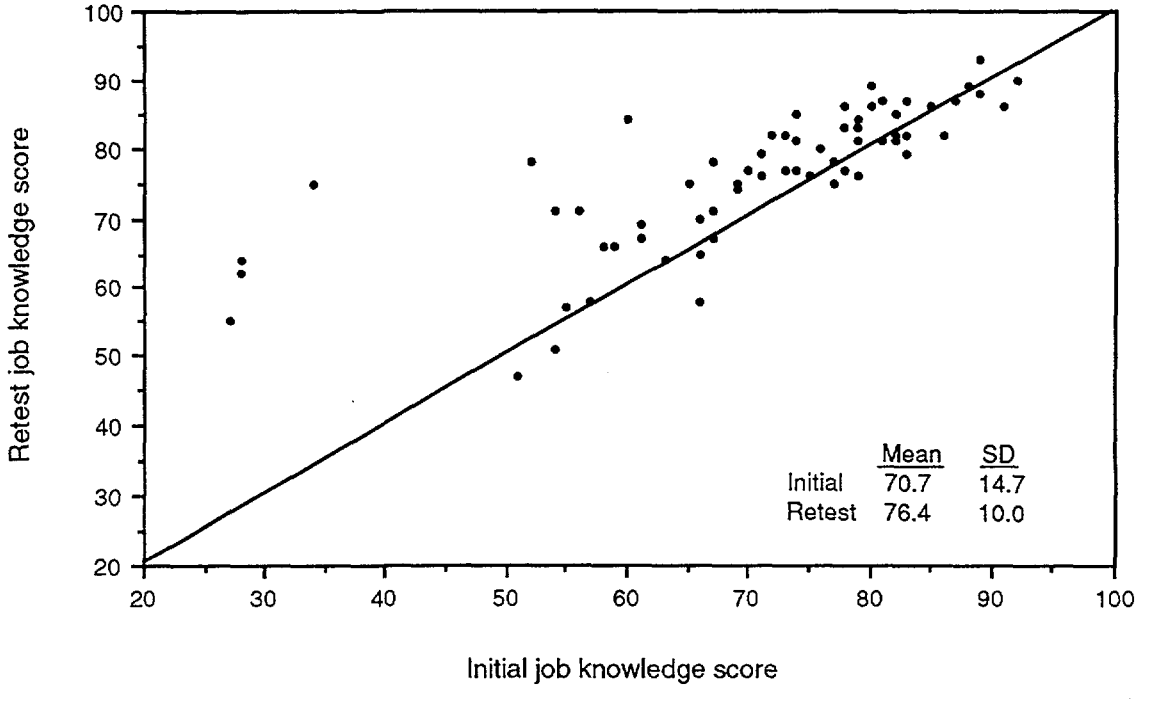


Figure 4. Scatterplot of initial and retest job knowledge scores for CH-46 mechanics (6112)

several important areas that may have had problems. For example, did administrators consistently agree on all tasks, or did some tasks present more scoring problems than others? Was a particular test administrator consistently deviant in scoring agreement relative to the other test administrators?

Figure 5 plots the percent agreement between test administrators for the 20 tasks of the automotive mechanics hands-on test. Agreement statistics are noted separately for the two testing locations (Camps Lejeune and Pendleton) to determine if similar trends were found at both sites. Corresponding figures for the four helicopter MOSs are presented in appendix B. Figure 5 shows that the agreement rates across tasks were consistently high and tracked reasonably well by site, although agreement indexes for the Pendleton test site were slightly higher. That is, tasks that experienced relatively high agreement rates at one site were likewise high at the other site, and vice versa. The lowest agreement index was noted at the Lejeune test site for task 7a, troubleshoot excessive oil consumption; agreement was found on only 87 percent of all steps scored. This task also had the second lowest agreement index at the Pendleton test site.

Figure 6 is a similar plot of step-level analysis noting the percent agreement between the 13 test administrators at the 2 testing sites. Similar plots for the helicopter MOSs are noted in appendix B. The Lejeune test site had two test administrators who consistently had low agreement indexes relative to other scorers: test administrator number 3 (85 percent) and number 7 (83 percent). In the other extreme, the Pendleton test site had one test administrator who had perfect agreement (100 percent). Despite these few outlying cases, most test administrators consistently agreed on over 90 percent of all steps scored.

Variance Components for Hands-on Performance Scores

Three hierarchical models were considered that progressively examined the complex structure of hands-on performance measurement. The first model, called the full model, estimated the variance attributable to a number of tangential factors that could affect hands-on scores. These tangential factors included a location variable to test for differences in performance across test sites and a time block variable to account for possible increases in scores over time as the result of test content becoming widely known. Based on the outcomes of these analyses, refinements to the model were made to examine more closely specific aspects of the measurement process; this was called the reduced model. A final model was examined that focused on a small sample of mechanics who were retested; this was called the retest model. Each of these models is described below in greater detail, characterizing the data structure and noting the differences between the samples for the automotive and helicopter specialties.

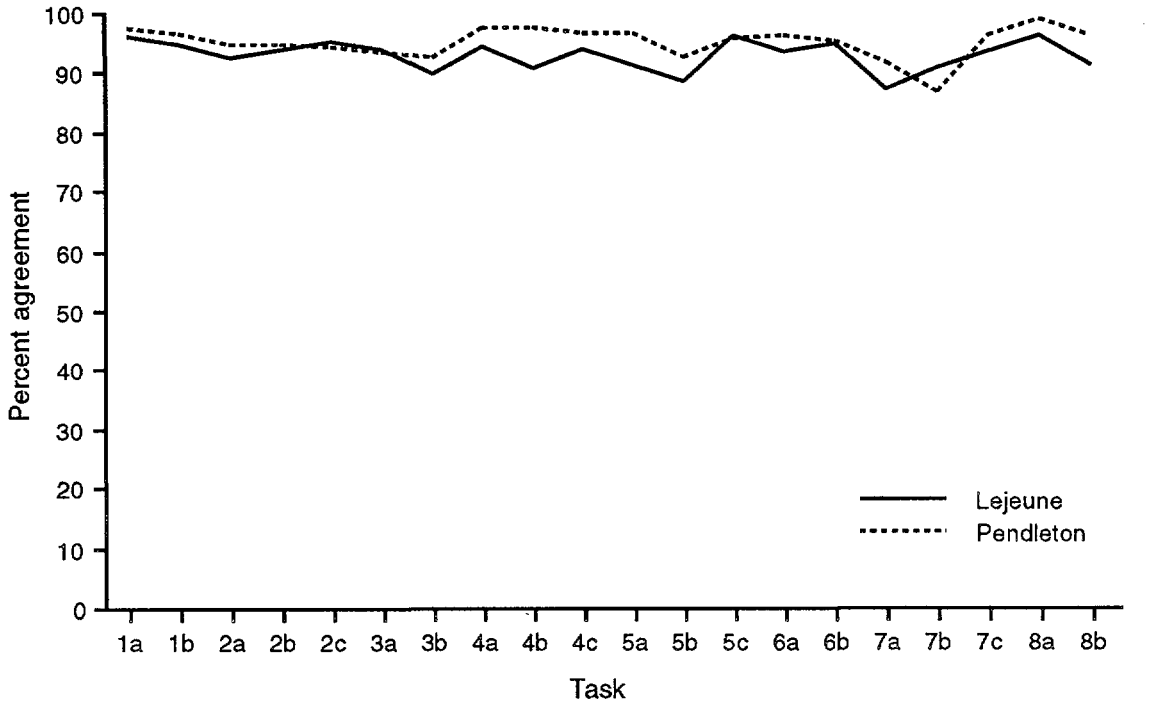


Figure 5. Test administrator agreement by task and site for ground mechanics (3521)

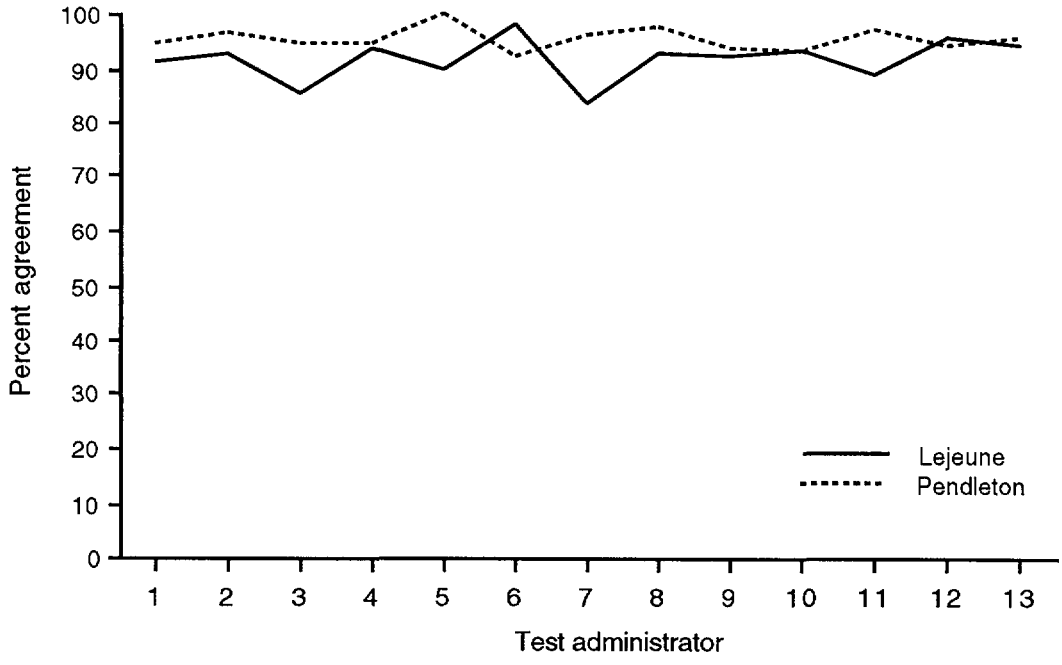


Figure 6. Test administrator agreement by administrator and site for ground mechanics (3521)

Full Measurement Model

The full measurement model included several extraneous factors that may affect hands-on scores but really should not. These factors were testing site and time block of test administration (Marines were divided into thirds based on when they were tested during the five-month testing period). Three other factors were also included in the model: task, test administrator, and Marine. The expectation was that variance in hands-on scores would primarily be a function of task difficulty and differences in individual performance levels (and their interaction), and that the variance components associated with test administrator, testing site, and block (and their interactions) would be negligible. The purpose of the full model was to test the effect of each factor and their interactions.

For the automotive mechanics, the full measurement model was fitted to 360 Marines who were scored by two test administrators over the course of the five-month testing period. Testing took place at two locations. Two different teams of test administrators conducted the testing, so test administrators are said to be "nested" within site. Artificial testing blocks of approximately equal size within test site were created to reflect a first, middle, or last chronological period of testing. For this model, Marines were tested only once, so Marines were nested within block and site. Although the hands-on test was composed of 18 tasks,[1] not all tasks were scored by two test administrators (shadow scored) for each Marine, nor were the same tasks necessarily shadow scored across Marines. However, all 18 tasks were performed and shadow scored for sufficient numbers of Marines during each testing block. To account for these differences in numbers of tasks across Marines and differences due to limitations of current statistical software to process unbalanced designs, tasks were randomly deleted to create a balanced design so that all Marines had performance scores for eight tasks that had been shadow scored.[2] Table 3 outlines the 15 combinations of these measurement factors and their interactions that potentially contribute to the variance of hands-on scores. An explanation is provided for interpreting large variance components associated with each factor.

---

1. The automotive hands-on test was composed of 20 tasks, but only 18 tasks required scoring by a test administrator. The other two tasks were scored according to an answer key and measured completion of forms to document maintenance actions and use of technical manuals.
2. More tasks could have been required as the common base for comparison but at the loss of sample size. Conversely, fewer tasks could have served as the base with larger sample sizes. Eight tasks provided a compromise between maximizing the number of tasks versus obtaining a reasonable sample of Marines. Analyses of data based on six and ten tasks resulted in similar conclusions.

Table 3. Factors for the full measurement model

Factor	Explanation for variance in hands-on scores
Site (S)	Sites differ in performance
Block (B)	Blocks differ in performance
Task (T)	Tasks differ in difficulty
Administrator (A:S)	Administrators differ in scoring
Marine (M:SB)	Marines differ in performance
S*B	Sites differ in performance by blocks
S*T	Task difficulty differs by site
B*T	Task difficulty differs over blocks
B*A:S	Administrators score differently over blocks
T*A:S	Administrators score tasks differently
T*M:SB	Marines differentially perform on tasks
A:S*M:SB	Administrators differentially score Marines
S*B*T	Task difficulty changes by site and block
B*T*A:S	Administrators differentially score tasks over blocks
T*A:S*M:SB, e	Random error

The full measurement model for the four helicopter MOSs included the same factors as for the automotive mechanics, except for the block factor. Hands-on testing was conducted sequentially for each aircraft and was typically completed in less than four weeks at each site so no blocking factor was created. Testing for the CH-53A/D helicopter (MOS 6113) was conducted at only one location because not enough mechanics were at the other. Fewer test administrators were available for shadow scoring of the helicopter testing; therefore, Marines who had at least three tasks shadow scored were included in the analysis. Tasks were randomly deleted for individuals with more than three tasks. Sample sizes ranged from 74 to 104 across the four aircraft. Again, table 3 will be useful for interpreting the results of variance component analyses; simply ignore any factor that includes the block variable (B).

Each hands-on test differs with respect to the variance of scores, which in turn affects the magnitude of the variance components. No standard exists against which to judge the magnitude of variance components. Rather, variance components are judged relative to other variance components in the analysis and are expressed as a percentage of the

total variance. Variance components, by definition, cannot be negative because variances must be greater than or equal to zero. In practice, relatively small negative variance components may be set to zero [2]; very large negative values typically imply model misspecification.

The results for the automotive mechanics are presented in table 4 and show that:

- Site and block had little or no effect on hands-on scores.
- Test administrators did not differ in their scoring; they were consistent across time (B*A:S), across tasks (T*A:S), and, most importantly, across Marines (A:S*M:SB).
- Tasks were found to differ in difficulty (T), and Marines differed in their level of job proficiency (M:SB).
- Over 60 percent of the variance in hands-on scores was attributable to Marines performing differently on different tasks (T*M:SB); that is, some mechanics performed better on some tasks but not on others--few have either mastered all tasks or perform poorly on all tasks.[1]

The results for the automotive mechanics indicate that test administrators did an excellent job of consistently scoring the performance they observed and that block and site factors were not required in the model to explain the variance in hands-on scores.

Table 5 presents the results for the three helicopter MOSs. The primary finding was that a substantial proportion of the variance in hands-on performance was attributable to factors including the site variable for all three aircraft. For the CH-46 mechanics (6112), in addition to large differences in scores due to site, site differences in task difficulty also contributed to the variance of hands-on scores (the S*T factor was 20 percent). Without further analysis to explain the site differences or adjustments to account for such differences (if they are simply an arbitrary product of the measurement process), the

---

1. The considerable variance attributable to the task by Marine factor should not be interpreted as contradictory to the high alpha coefficient estimate of reliability. Alpha coefficients are primarily influenced by inter-item correlations and do not necessarily reflect differences in mean item performance. Variance components are totally a function of mean differences attributable to a specific source.

aggregation of scores across sites is not warranted.[1] As with the automotive mechanics, most of the variance in helicopter hands-on scores was caused by the interaction between Marines and tasks. The variance component for the interaction of Marines and test administrators was consistent across the three MOSs at about 10 percent, which was somewhat larger than desired, but the magnitude of these estimates may be partially confounded by the substantial site effect.

Table 4. Variance component estimates for full measurement model for automotive mechanics

Measurement factor	Variance component	Percentage
Site (S)	0.3	0.1
Block (B)	-9.5	0.0
Task (T)	62.7	9.8
Administrator (A:S)	1.4	0.2
Marine (M:SB)	45.1	7.0
S*B	22.3	3.5
S*T	3.3	0.5
B*T	29.1	4.5
B*A:S	2.8	0.4
T*A:S	15.1	2.4
T*M:SB	412.7	64.4
A:S*M:SB	-9.9	0.0
S*B*T	10.6	1.7
B*T*A:S	28.9	4.5
T*A:S*M:SB, e	6.4	1.0

NOTE: Marine and administrator are nested within site. Percentages are based on respective variance components divided by the total of all positive variance components.

1. Further analyses were undertaken to examine specifically the magnitude and possible cause for the consistent effect of site on the hands-on scores for helicopter mechanics. These analyses and the adjustments subsequently applied to account for the site differences are noted elsewhere [1].

Table 5. Variance component estimates for full measurement model for helicopter mechanic specialties

Measurement factor	Variance component	Percent-age	Variance component	Percent-age	Variance component	Percent-age
Site (S)	107.9	(21.3)	214.3	(25.2)	105.9	(27.8)
Task (T)	5.6	(1.1)	127.7	(15.0)	7.2	(1.9)
Marine (M:S)	16.6	(3.3)	34.5	(4.1)	4.6	(1.2)
Administrator (A:S)	3.2	(0.6)	29.2	(3.4)	-8.0	(0.0)
S*T	101.6	(20.0)	-26.9	(0.0)	27.6	(7.2)
T*M:S	218.5	(43.0)	354.4	(41.6)	161.2	(42.3)
T*A:S	-0.6	(0.0)	5.0	(0.6)	22.3	(5.9)
M:S*A:S	54.3	(10.7)	86.5	(10.2)	52.4	(13.8)
S*T*A:S*M:S, e	-45.6	(0.0)	-99.5	(0.0)	-55.8	(0.0)

NOTE: Both Marines and administrators are nested within site. Percentages are based on respective variance components divided by the total of all positive variance components.

Reduced Measurement Model

Results from the full measurement model for the automotive mechanics indicated that the site and block variables did not affect the variance of the hands-on scores. The results for the helicopter mechanics found substantial site effects on hands-on scores. The models for all MOSs were reduced in terms of the number of factors based on these outcomes to more closely address aspects of the measurement process related to test administrators. Table 6 details the factors for which variance components were explicitly estimated.

Table 6. Factors for the reduced model

Factor	Explanation for variance in hands-on scores
Task (T)	Tasks differ in difficulty
Administrator (A)	Administrators differ in scoring
Marine (M)	Marines differ in job performance scores
T*A	Administrators differentially score tasks
T*M	Marines differentially perform on tasks
A*M	Administrators differentially score Marines
T*A*M, e	Random error

Table 7 presents the variance component estimates for the reduced model for the automotive mechanics. Again, test administrators were found to cause minimal variance in the hands-on scores. Although slightly over 6 percent of the hands-on variance was a function of test administrators applying different scoring standards across tasks (T*A), test administrators were consistent in how they scored different individuals (A*M). The bulk of the variance in hands-on scores was still a function of the task by Marine interaction (T*M), differences in task difficulty (T), and individual differences (M).

Table 7. Variance component estimates for reduced measurement model for automotive mechanics

Measurement factor	Variance component	Percentage
Task (T)	76.5	12.1
Administrator (A)	6.6	1.0
Marine (M)	57.6	9.1
T*A	40.3	6.4
T*M	441.9	69.8
A*M	-12.2	0.0
T*A*M, e	9.8	1.6

NOTE: Percentages are based on respective variance components divided by the total of all positive variance components.

The same reduced model was estimated for each of the helicopter MOSs separately by test site because of the noted site effects. Sample sizes for each analysis were 40 or greater (only 11 MOS 6112 mechanics at MCAS Tustin were shadow scored and, therefore, deleted from the analysis). Some inconsistencies in the magnitude of the variance component estimates were noted across sites that possibly were also a function of small sample sizes, but the general trends were quite similar to the findings for the automotive mechanics (see table 8):

- The largest variance component was the interaction between task and Marine (T*M), followed by the components for task (T) and Marine (M).

- Effects due to test administrator tended to be small, although there were some exceptions (T*A = 20.8 percent for MOS 6115 at New River, and A*M = 10.5 percent for MOS 6115 at Tustin).

In summary, no substantial effects due to test administrators were found for the automotive and helicopter specialties.

Retest Measurement Model

The previous two models have demonstrated that test administrators can reliably score hands-on performance across test content, across individual examinees, and over time. Differences in the performance due to examinees were somewhat smaller than expected; however, each data set was based on highly selected samples that restricted the true variance due to individual differences. A third model was examined to explore specifically the performance levels of individual mechanics and to what extent such performance improves on retesting. In addition, the consistency of test administrators scoring the same person's performance over time was examined.

Strict data collection plans were devised to ensure that the retested Marines who were shadow scored were tested by the same pair of test administrators on both occasions for exactly the same hands-on content. Relatively few Marines were retested and, of these, even fewer were shadow scored. A total of 47 automotive mechanics and 15 CH-46 mechanics satisfied both requirements. To examine more closely the effects of test administrator on hands-on scores, the same pair of test administrators scored a Marine's performance on the first half of the hands-on test and another pair scored the second half of the test. The data from these two halves of the test were then analyzed as if they were replications for the same person (i.e., some tasks differences were potentially ignored and the task factor may not accurately reflect the total variance due to task differences). Table 9 details the other factors of the retest measurement model.

Table 8. Variance component estimates for reduced measurement model for helicopter mechanic specialties

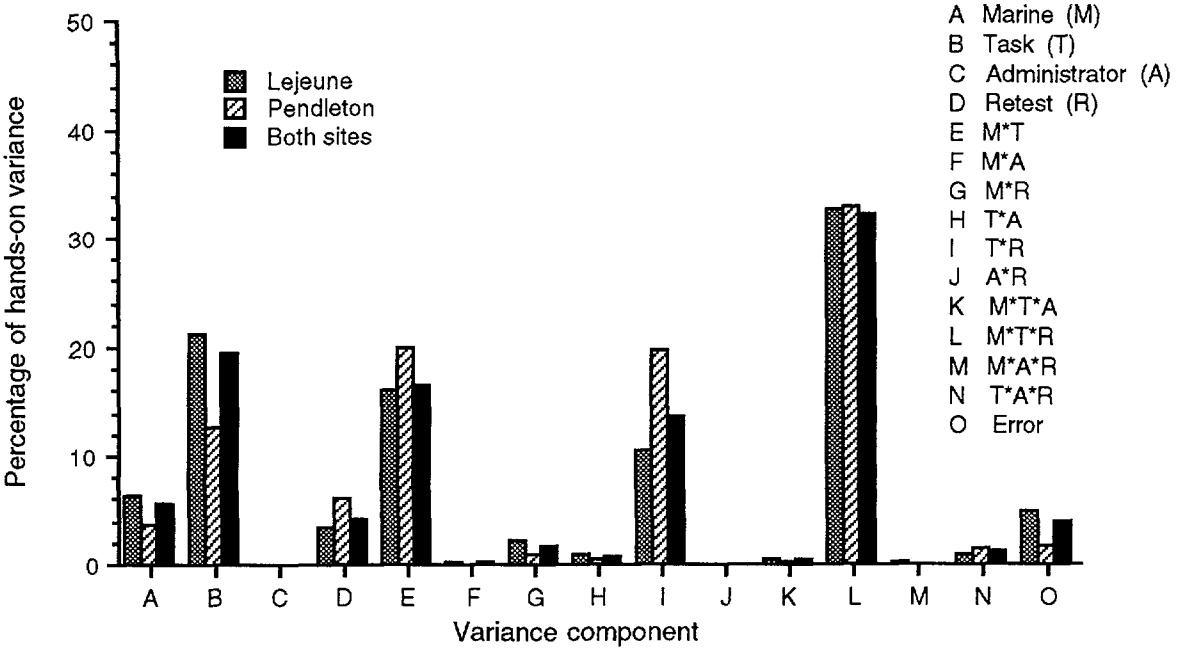
	6112	6113	6114		6115	
Measurement factor	NR	TS	NR	PN	NR	TS
Task (T)						
Variance component	113.9	6.8	36.1	112.3	22.6	82.9
Percentage	35.3	1.5	15.2	12.4	19.0	19.7
Administrator (A)						
Variance component	1.7	15.0	20.8	34.5	-4.6	23.9
Percentage	0.5	3.2	8.8	3.8	0.0	5.7
Marine (M)						
Variance component	37.2	36.7	9.4	87.2	7.5	18.0
Percentage	11.5	7.8	4.0	9.6	6.3	4.3
T*A						
Variance component	0.5	20.3	15.4	-7.4	24.7	-16.6
Percentage	0.2	4.3	6.5	0.0	20.8	0.0
T*M						
Variance component	145.5	362.0	142.0	625.6	58.3	251.2
Percentage	45.1	76.8	59.9	69.1	48.9	59.8
A*M						
Variance component	-13.1	30.8	13.3	46.5	-10.2	44.3
Percentage	0.0	6.5	5.6	5.1	0.0	10.5
T*A*M,e						
Variance component	23.5	-38.6	-18.5	-63.3	6.0	-47.0
Percentage	7.3	0.0	0.0	0.0	5.0	0.0
Sample size	63	78	55	49	50	40

NOTE: Variance components and percentage of total variance are reported separately by site: New River (NR), Tustin (TS), and Pendleton (PN). Percentages are based on respective variance components divided by the total of all positive variance components.

Table 9. Factors for the retest model

Factor	Explanation for variance in hands-on scores
Marine (M)	Marines differ in job performance scores
Task (T)	Tasks differ in difficulty
Administrator (A)	Administrators differ in scoring
Retest (R)	Performance changes on retesting
M*T	Marines differentially perform on tasks
M*A	Administrators differentially score Marines
M*R	Marines differentially improve on retesting
T*A	Administrators differentially score tasks
T*R	Task difficulty changes on retesting
A*R	Administrators score differently on retesting
M*T*A	Administrators differentially score Marines on different tasks
M*T*R	Marines differentially improve on different tasks on retesting
M*A*R	Administrators differentially score Marines on retesting
T*A*R	Administrators differentially score tasks on retesting
Error	Random error

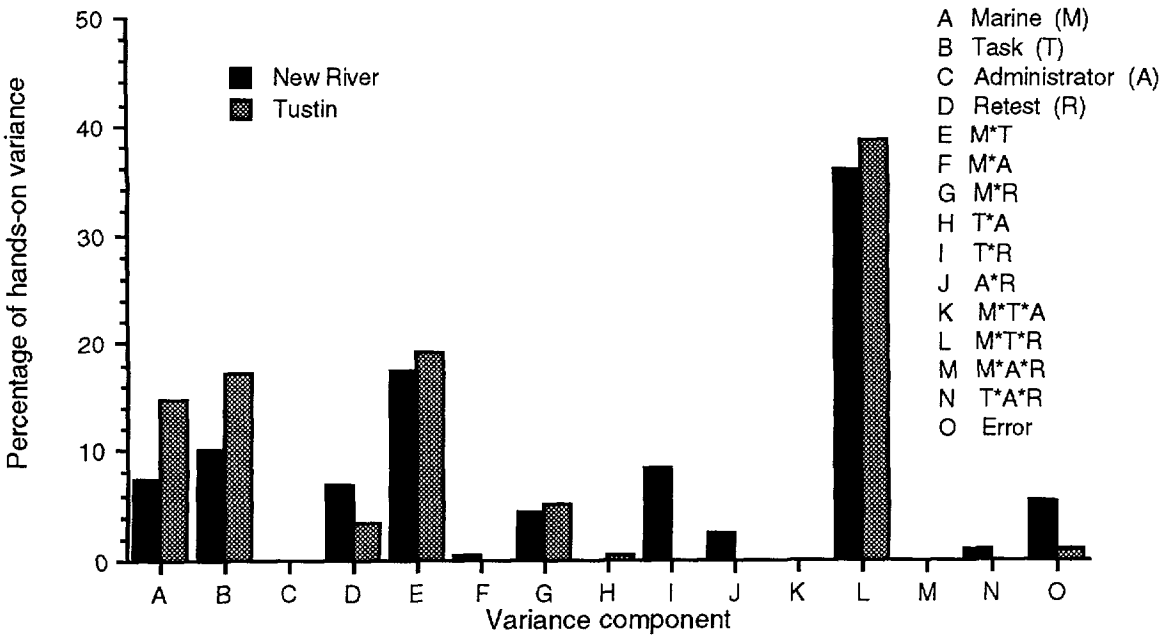
Figure 7 plots the percentage of hands-on variance explained by each factor of the retest model for the automotive mechanics. Variance components were estimated separately for each site because this variable was not explicitly modeled. The results across sites were essentially the same, consistent with the lack of a site effect for the earlier automotive mechanic data set. All seven variance components involving the test administrator were negligible--test administrators were consistent in scoring across individual Marines (M*A), tasks (T*A), retesting (A*R), Marines on different tasks (M*T*A), Marines on retesting (M*A*R), and different tasks on retesting (T*A*R). By far, the largest variance component was the interaction of Marine, task, and retest--Marines differentially improved on different tasks on retesting. The magnitude of this component may partially reflect motivational differences of examinees to perform each task well on retesting. Consistent with the findings of the previous models for automotive mechanics, the task factor and the interaction of Marine and task were substantial contributors to the variance of hands-on scores. Mean task performance does improve on retesting after 10 to 14 days (T*R). The degree of intervening practice between the two test sessions was not known. However, such an outcome has implications for the potential effectiveness and impact of training to enhance mechanical job performance.



NOTE: For variance components where no bar is indicated, the percentage of total variance is equal to zero.

Figure 7. Variance-component percentages for retest model: automotive mechanics (3521)

Unlike the automotive mechanic data, test administrators for the CH-46 worked in pairs to score either 11 tasks or 8 tasks (difference is the result of the configuration of tasks into the eight testing stations). To achieve a balanced design of 8 tasks for each mechanic and pair of administrators, 3 tasks were randomly deleted for examinees who were scored by two administrators on 11 tasks. Figure 8 shows the same variance component estimates for the CH-46 helicopter. Caution should be used in interpreting these outcomes because they are based on relatively few mechanics per site (15 mechanics total). Despite the small sample size, the results for the CH-46 mechanics were quite consistent for both sites as well as similar to the findings for the automotive mechanics. Again, all variance components involving administrators were nonexistent or small, and the three largest variance components were the same as automotive mechanics. The results from these retest analyses closely paralleled the findings of the previous two models.



NOTE: For variance components where no bar is indicated, the percentage of total variance is equal to zero.

Figure 8. Variance-component percentage for retest model: CH-46 mechanics (6112)

CONCLUSIONS

The assessment of performance via hands-on tests is potentially replete with factors that can degrade their reliability. The quality and consistency of scoring by test administrators is a primary concern. The analyses of this research memorandum have focused on estimating the reliability of the hands-on tests (and job knowledge tests) for several mechanical specialties so that policy-makers can understand the factors that affect such performance measurements. Reliability estimates focused not only on the test administrators but also on the stability of individual performance, the difficulty of mechanical tasks, differences in scoring and performance across test sites, and how these factors change over time.

Several different methods of estimating reliability found the same results for the hands-on and job knowledge tests--both performance tests resulted in consistent and reliable measurements. Analyses of the factors that contribute to the variance of hands-on scores showed that the test administrators were interchangeable and had little impact on the overall hands-on performance scores. These findings were probably the result of several strategies taken during the test development and administration process to enhance the reliability of measurements:

- Job knowledge tests were thoroughly pretested to identify problem items or ambiguous response alternatives for revision or deletion. Job knowledge tests were of sufficient length to provide detailed subtest information and to enhance the possibilities of obtaining acceptable test reliability estimates.

- The development process for hands-on tests focused on identifying scorable steps that required performance of specific observable actions. There were few, if any, tasks that presented test administrators with ambiguous performance requirements that would make the task difficult to score.

- The hands-on performance tests were administered by retired or former Marines with relevant job experience who had no vested interest in the outcomes of the testing. Such reliable scoring results may not have been obtained by active duty personnel scoring the performance of their subordinates or in situations in which only cursory training of test administrators was provided.

- Training of test administrators was comprehensive to ensure that they understand not only how to perform all tasks but how to administer the tasks in a standardized manner. Attention was devoted to what types of feedback were appropriate to questions by the mechanics being tested.

- The continual monitoring of test administrator scoring over the testing period (by entering the data daily into a computer) allowed scoring problems to be identified immediately and resolved. Test administrators were debriefed on their scoring patterns and could identify performance steps or scoring criteria that required further discussion, training, or modification.

Despite these efforts to obtain reliable measurements, hands-on test scores were found to vary by testing site for all of the helicopter MOSs. Detailed training materials were prepared and the same team of trainers was used at both locations to standardize conditions in an attempt to minimize such errors across sites. The magnitude of the site effect was consistent across all aircraft (except for the CH-53A/D, which was tested at only one site). It is interesting that a site effect was not found for the testing of automotive mechanics where essentially the same training and monitoring strategies were used. The presence of the site effect suggests that additional monitoring procedures are needed to check scoring consistency not only within a test site but also across test sites.

Three different measurement models were examined that resulted in essentially the same findings pertaining to the largest sources of variance in hands-on scores: Marines differentially performing on tasks (M*T), tasks differing in difficulty (T), and individual performance differences (M). For all three models, factors involving test administrators contributed relatively little variance to hands-on scores. Test administrators were found to score performance consistently across examinees, test content, and time.

The hands-on performance tests were composed of 15 to 20 tasks that were sampled to represent the extremely large domain of mechanical job requirements. The large variance component for the task factor of each MOS indicates that sampling error associated with selection of tasks included in the hands-on test may have significant implications for the interpretation of hands-on performance scores. Given such extensive job requirements for Marine Corps mechanics, certain aspects of the job domain may not be represented if the tasks are not properly sampled or too few tasks are tested. It follows that explicit efforts should be made to sample tasks that are representative of the full range of job requirements. Likewise, the tests should focus on the critical aspects of the task performance so that redundant behaviors are not measured multiple times across tasks. More discrete tasks should be administered at the expense of administering fewer tasks that probably contain highly correlated steps.

The substantial differences found for Marine mechanics performing differently across tasks (the M*T and M*T*R interactions) implies that there is not generalized mechanical expertise. In other words, some mechanics perform well on some tasks and other mechanics perform well on other tasks, but few perform well or poorly on all tasks. Some, but not all, of the performance differences may possibly be explained by individual differences in aptitude. Some variance in performance is possibly attributable to differences in experiences the mechanics have received from various training programs or on-the-job instruction.

The magnitude of the variance components involving the retest factor indicates that task performance is responsive to practice effects. Such improvements in performance could even be increased if such practice were refined and presented in a focused training context. However, the Marine-by-retest factor (Marines differentially improve on retesting) implies that everyone may not respond equally well to the same training. Instructional courses or on-the-job training may have to take multiple approaches to teaching mechanical content in attempting to improve individual performance levels. Likewise, further examination of the task-by-retest interaction may assist in determining which tasks are most responsive to practice. The JPM Project also collected task-level information regarding the recency and frequency of task performance to be able to address questions of the perishability of mechanical skills and possibly to assist in the timing of individual and unit training before unit deployments.

The Marine Corps JPM Project has developed reliable measures of hands-on performance and job knowledge for several mechanical maintenance specialties. Confidence can be placed in the interpretation of the resulting scores, and conclusions of further analyses will not be affected by the instability or inconsistency of the performance scores. The reliability analyses also point to several potential implications for the training community. These implications will be addressed more completely in later analyses of the mechanical performance data.

THIS PAGE INTENTIONALLY LEFT BLANK

REFERENCES

[1] CNA Research Memorandum 91-242, *Development and Scoring of Hands-on Performance Tests for Mechanical Maintenance Specialties*, by Neil B. Carey and Paul W. Mayberry, forthcoming

[2] R. L. Brennan. *Elements of Generalizability Theory*. Iowa City, IA: American College Testing Publications, 1983

APPENDIX A

COMPUTATION OF RELIABILITY ESTIMATES CORRECTED FOR RANGE RESTRICTION

APPENDIX A

COMPUTATION OF RELIABILITY ESTIMATES CORRECTED FOR RANGE RESTRICTION

Estimates of reliability, like validity coefficients, are affected by range restriction due to the selection process. An estimate of the population reliability coefficient ($\hat{\rho}_{xx}$) can be computed as follows:

$$\hat{\rho}_{xx} = 1 - \frac{s_x^2}{\sigma_x^2}(1 - r_{xx}) \quad , \qquad \text{(A-1)}$$

where s_x^2 and σ_x^2 are the sample and population variances, respectively, and r_{xx} is the sample reliability. Equation A-1 assumes that the error variances are equal for both the sample and population. Given that mechanical performance measures have no ceiling or floor effects, this assumption should be satisfied.

Table A-1 provides the sample and population standard deviations for the hands-on and job knowledge tests. Population standard deviations were obtained from the range correction algorithm that accounts for selection effects on all Armed Services Vocational Aptitude Battery (ASVAB) subtests [A-1, A-2]. The sample reliabilities are presented in table A-2. Computed population reliabilities are reported in table 1 of the text.

Table A-1. Sample and population standard deviations for hands-on performance tests and job knowledge tests

	MOS				
	3521	6112	6113	6114	6115
Hands-on performance test					
Sample	8.33	7.26	8.77	7.64	7.18
Population	9.94	9.12	10.92	9.06	7.99
Job knowledge test					
Sample	11.40	13.12	10.79	9.73	9.83
Population	13.74	17.41	14.83	12.42	10.61

Table A-2. Sample reliability estimates for hands-on performance tests and job knowledge tests

	MOS				
Reliability measure	3521	6112	6113	6114	6115
Hands-on performance test					
Test-retest	.70	.81			
Split-halves	.71	.80	.85	.74	.84
Alpha coefficient	.73	.81	.81	.69	.77
Scorer agreement	.94	.97	.94	.94	.93
Job knowledge test					
Test-retest	.61	.77			
Split-halves	.91	.95	.93	.92	.91
Alpha coefficient	.90	.95	.92	.90	.92

REFERENCES

[A-1] H. Gulliksen. *Theory of Mental Tests*. New York: Wiley, 1950

[A-2] CNA Research Contribution 336, *A Method To Correct Correlation Coefficients for the Effects of Multiple Curtailment*, by Thomas L. Mifflin and Stephen M. Verna, Aug 1977

APPENDIX B

TEST ADMINISTRATOR AGREEMENT INDEXES FOR HELICOPTER SPECIALTIES

APPENDIX B

TEST ADMINISTRATOR AGREEMENT INDEXES FOR HELICOPTER SPECIALTIES

Step-level percent agreement indexes between test administrator pairs were computed for each of the four helicopter specialties, separately for each testing site. The agreement indexes were computed for each task and then for each test administrator to note any particular problem tasks or deviant test administrators. Figures B-1 through B-4 plot the results for the task comparisons for the four helicopter specialties, and figures B-5 through B-8 are the same plots across test administrators.

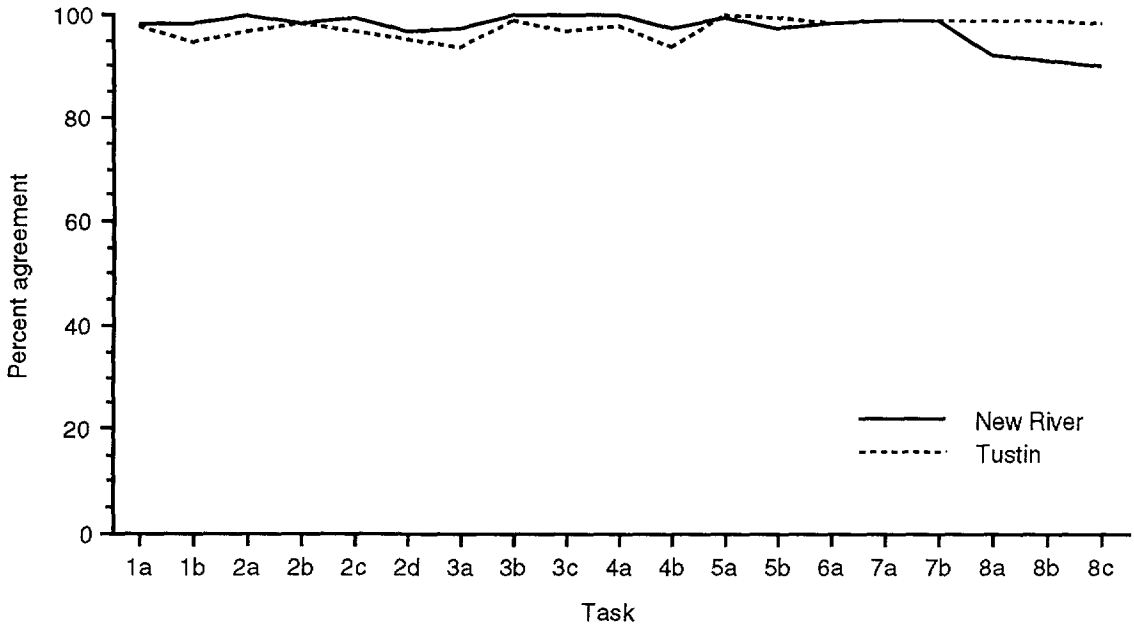


Figure B-1. Test administrator agreement by task and site for CH-46 mechanics (6112)

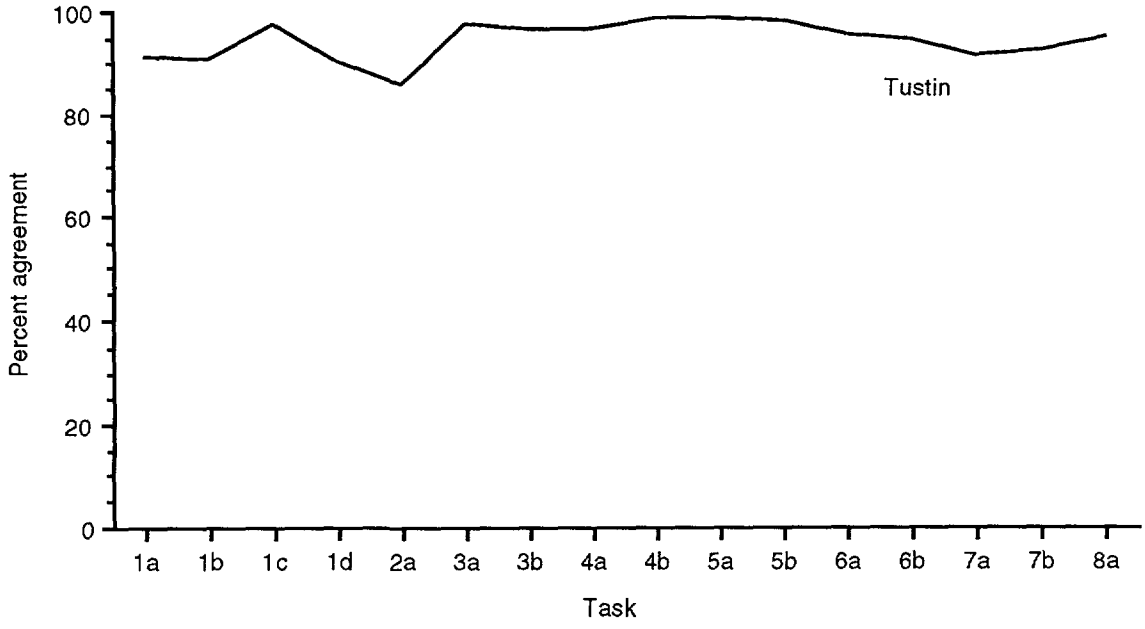


Figure B-2. Test administrator agreement by task for CH-53A/D mechanics (6113)

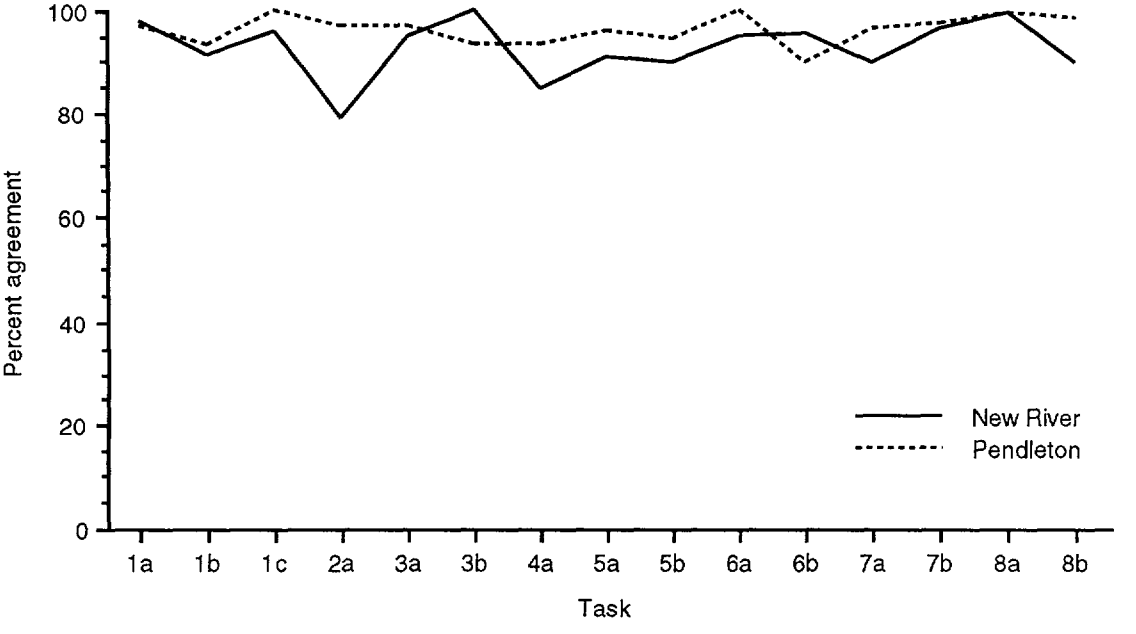


Figure B-3. Test administrator agreement by task and site for U/AH-1 mechanics (6114)

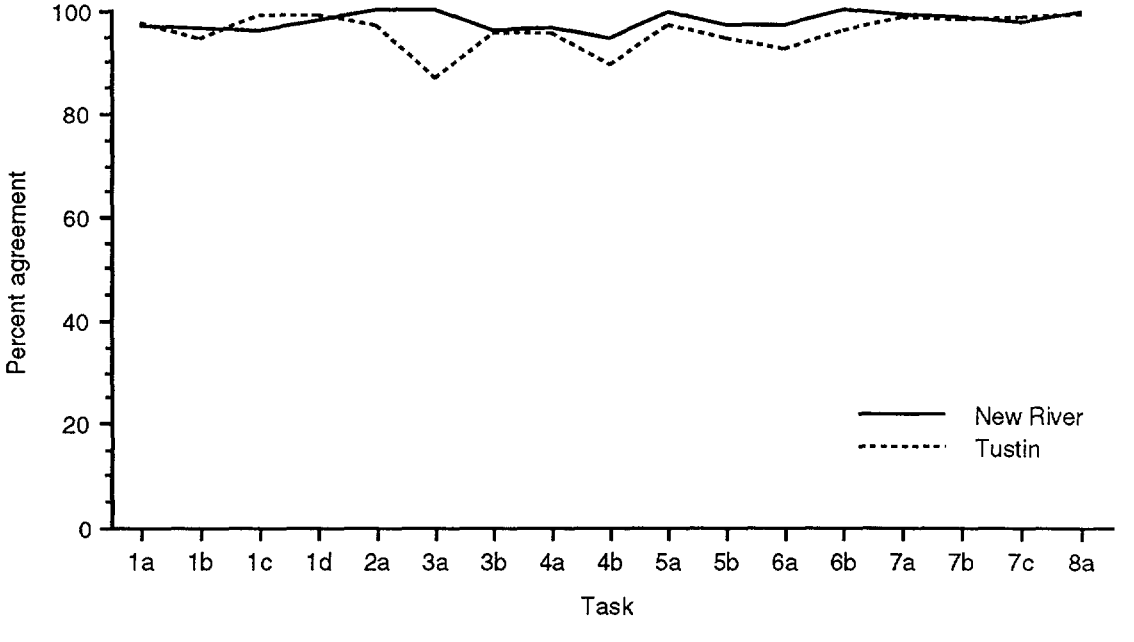


Figure B-4. Test administrator agreement by task and site for CH-53E mechanics (6115)

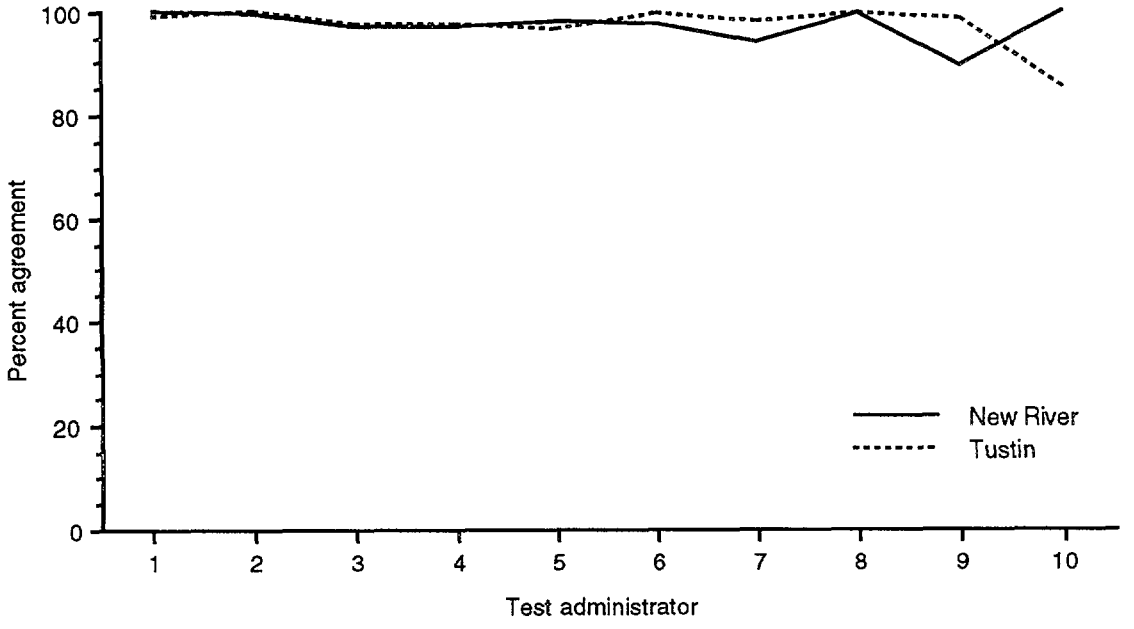


Figure B-5. Test administrator agreement by administrator and site for CH-46 mechanics (6112)

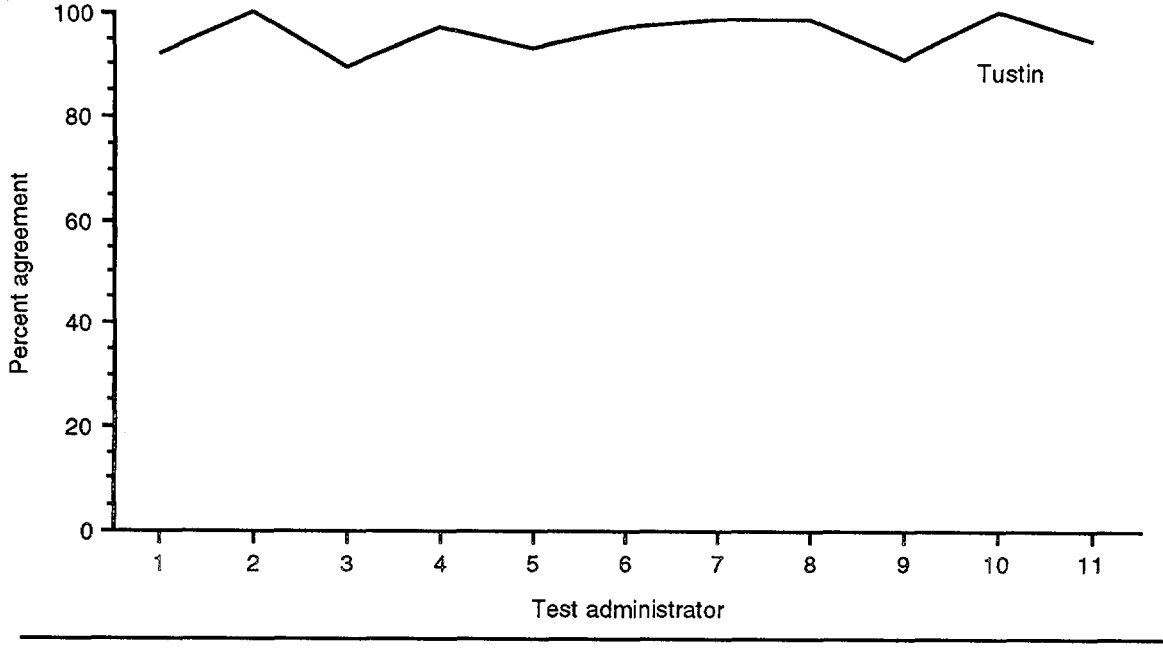


Figure B-6. Test administrator agreement by administrator and site for CH-53A/D mechanics (6113)

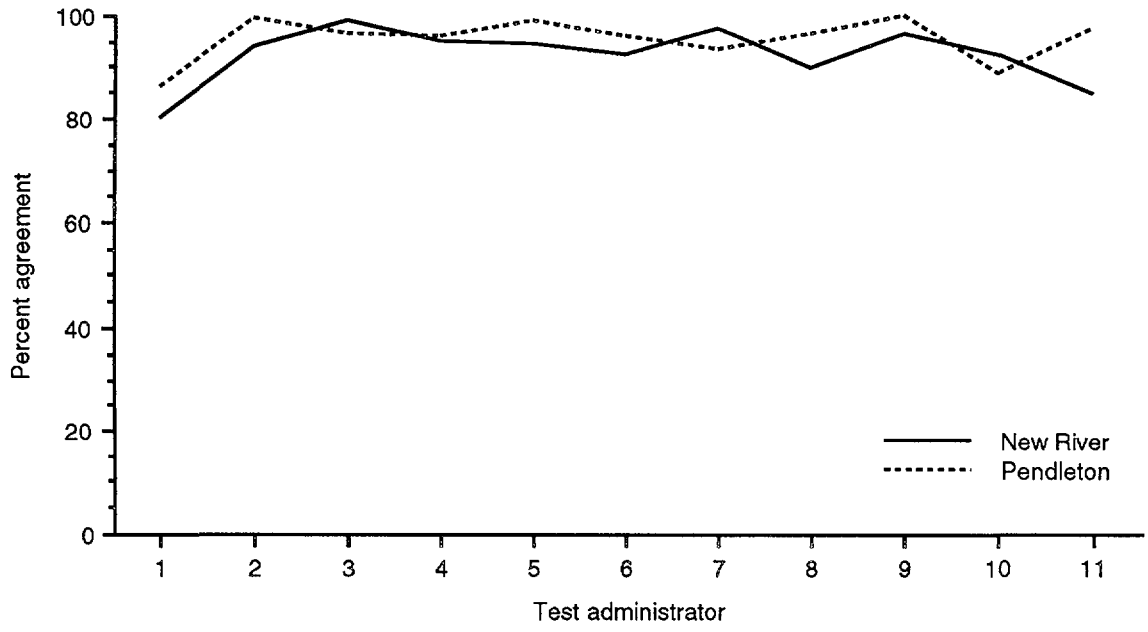


Figure B-7. Test administrator agreement by administrator and site for U/AH-1 mechanics (6114)

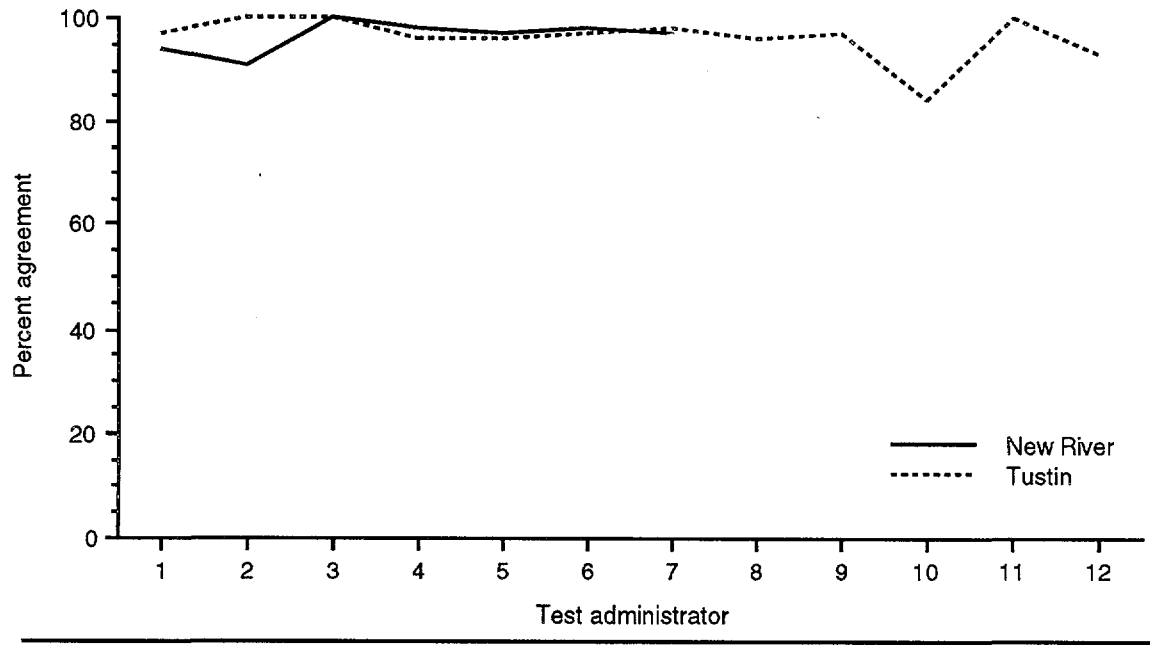


Figure B-8. Test administrator agreement by administrator and site for CH-53E mechanics (6115)